Coursera Statistical Inference Project

Erica Pehrsson

6/28/2018

Part 1

Overview

This report demonstrates an example of the Central Limit Theorem, which states that regardless of the structure of the underlying population distribution, averages of samples taken from that distribution will converge on the population mean. The standard error of the mean (the standard deviation of the sample means) will be approximately the standard deviation divided by the square root of the sample size. This report demonstrates the concordance between a large number of random draws from the exponential distribution and the theoretical values described above.

```
library(plyr)
library(ggplot2)
```

Simulations

The below code samples 40 random values from an exponential distribution with rate 0.2, repeated one thousand times. Then, it takes the mean of each sample of 40 values.

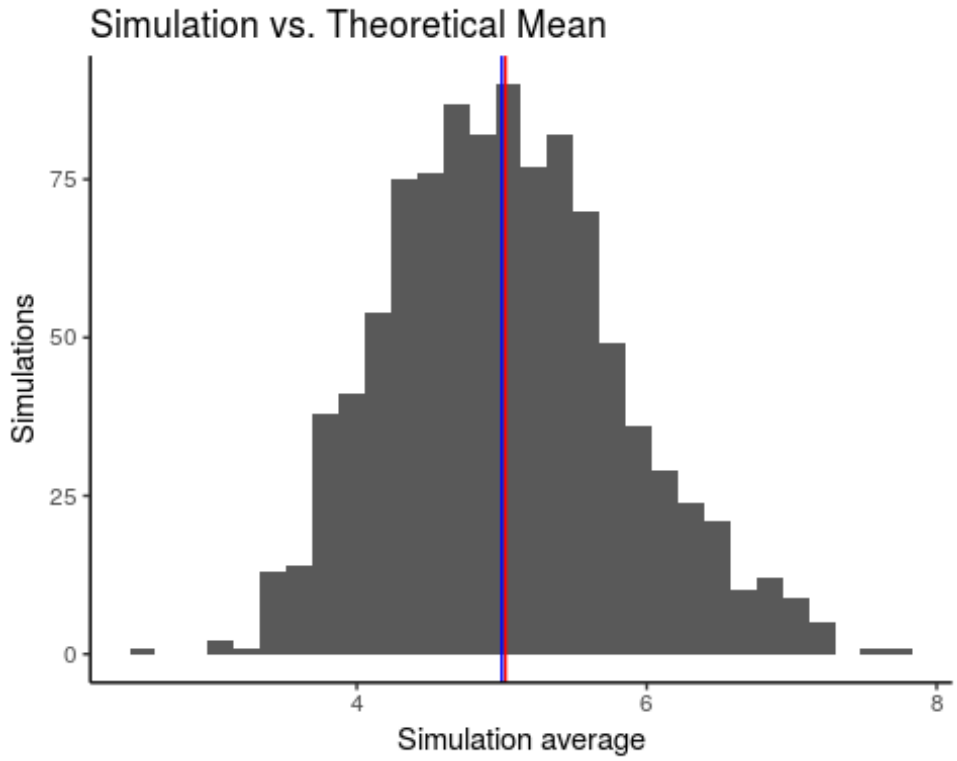
```
lambda=0.2
N=40
sim=1000
set.seed(4)

simulations = ldply(seq(1,sim,1),function(x) rexp(N,lambda))
simulations$Mean = apply(simulations,1,mean)
```

Sample Mean versus Theoretical Mean

```
# Sample mean
mean.sample = mean(simulations$Mean)

# Theoretical mean
mean.theory = 1/lambda
```

Here, I show the distribution of the sample means of 40 draws from an exponential distribution with rate 0.2. The mean of the sample distribution (red) is `mean.sample`, roughly the mean of the underyling population (blue), `mean.theory`.

```
ggplot(simulations,aes(x=Mean)) + geom_histogram() + geom_vline(xintercept=me
an.sample, color="red") + geom_vline(xintercept=mean.theory, color="blue") +
  theme_classic() + ggtitle("Simulation vs. Theoretical Mean") + ylab("Simula
tions") + xlab("Simulation average")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



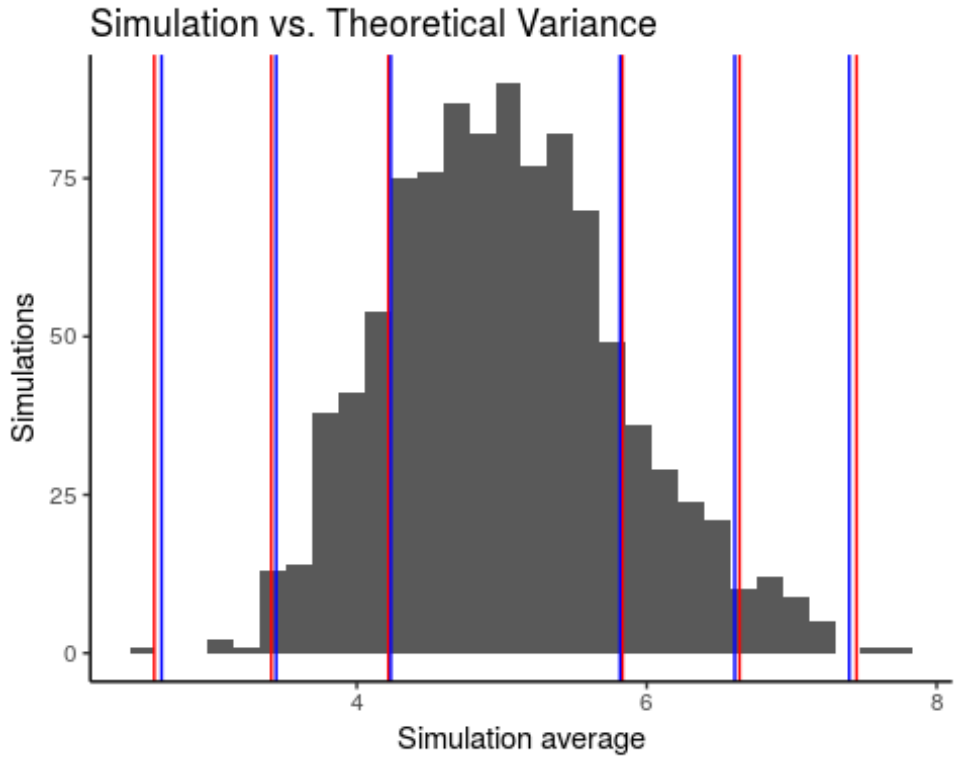
Sample Variance versus Theoretical Variance

```
# Sample variance
variance.sample = sd(simulations$Mean)^2
sd.sample = sqrt(variance.sample)

# Theoretical variance
variance.theory = (1/lambda)^2/N
sd.theory = sqrt(variance.theory)
```

Similarly, the variance of the observed sample distribution (`variance.sample`) is similar to the theoretical predicted variance (`variance.theory`). The histogram below shows 1, 2, and 3 standard deviations from the sample distribution mean (red), compared to the standard deviations predicted by the formula for standard error of the mean (blue).

```
ggplot(simulations,aes(x=Mean)) + geom_histogram() +
  geom_vline(xintercept=mean.sample+sd.sample, color="red") + geom_vline(xint
ercept=mean.sample-sd.sample, color="red") +
  geom_vline(xintercept=mean.sample+2*sd.sample, color="red") + geom_vline(xi
ntercept=mean.sample-2*sd.sample, color="red") +
```

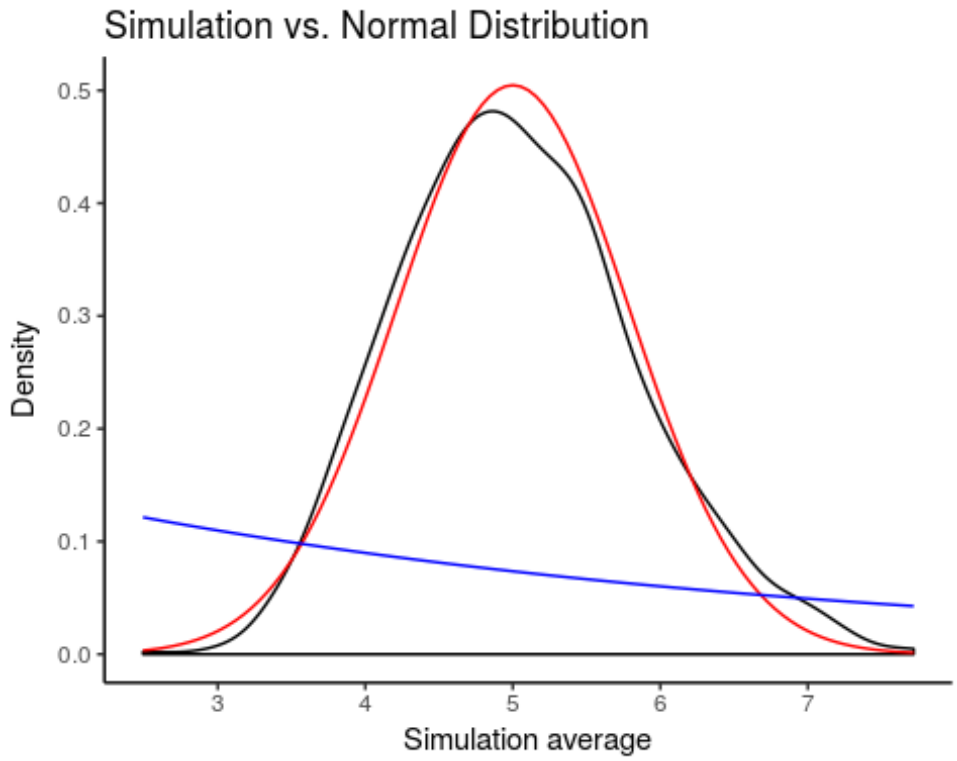
```
  geom_vline(xintercept=mean.sample+3*sd.sample, color="red") + geom_vline(xi
ntercept=mean.sample-3*sd.sample, color="red") +
  geom_vline(xintercept=mean.sample+sd.theory, color="blue") + geom_vline(xin
tercept=mean.sample-sd.theory, color="blue") +
  geom_vline(xintercept=mean.sample+2*sd.theory, color="blue") + geom_vline(x
intercept=mean.sample-2*sd.theory, color="blue") +
  geom_vline(xintercept=mean.sample+3*sd.theory, color="blue") + geom_vline(x
intercept=mean.sample-3*sd.theory, color="blue") +
  theme_classic() + ggtitle("Simulation vs. Theoretical Variance") + ylab("Si
mulations") + xlab("Simulation average")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Distribution

Finally, I compare the density of the sample distribution (black) to a normal population with mean identical to the underlying exponential distribution rate and standard deviation identical to the predicted standard error of the mean (red). There is high concordance between the two. The two distributions are very similar. This is in contrast to the exponential distribution (blue), which follows a much different pattern.

```
ggplot(simulations,aes(x=Mean)) + geom_density() +
  stat_function(fun=dnorm,args=list(mean = mean.theory, sd = sd.theory),color
="red") +
  stat_function(fun=dexp,args=list(rate = lambda),color="blue") +
  theme_classic() + ggtitle("Simulation vs. Normal Distribution") + ylab("Den
sity") + xlab("Simulation average")
```


Simulation vs. Normal Distribution
Density
0.5
0.4
0.3
0.2
0.1
0.0
3
4
5
6
7
Simulation average